TEC-0033

(2)

Passive Recovery of Scene Geometry for an Unmanned Ground Vehicle

Robert C. Bolles

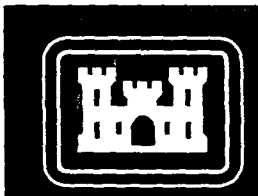
SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025-3493

May 1993

Approved for public release; distribution is unlimited.

Prepared for:
Advanced Research Projects Agency
1400 Wilson Boulevard
Arlington, VA 22209-2308

Monitored by:
U.S. Army Corps of Engineers
Topographic Engineering Center
Fort Belvoir, Virginia 22060-5546



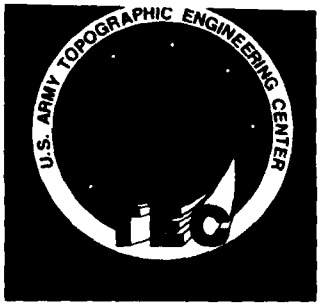
US Army Corps of Engineers
Topographic Engineering Center

T

E

C

34 pp



Destroy this report when no longer needed.
Do not return it to the originator.

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

The citation in this report of trade names of commercially available products does not constitute official endorsement or approval of the use of such products.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	May 1993	Annual Report Nov. 1991 - Nov. 1992

4. TITLE AND SUBTITLE

Passive Recovery of Scene Geometry for an Unmanned Ground Vehicle

5. FUNDING NUMBERS

DACA-76-92-C-0003
ARPA Order No: 8323

6. AUTHOR(S)

Robert C. Bolles

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

SRI International
333 Ravenswood Avenue
Menlo Park, CA 94025-3493

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Advanced Research Projects Agency
1400 Wilson Boulevard, Arlington, VA 22209-2308

U.S. Army Topographic Engineering Center
Fort Belvoir, VA 22060-5546

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

TEC-0033

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The primary goal of this project is to develop a passive technology for recovering scene geometry in order to support an Unmanned Ground Vehicle (UGV) operating in a general outdoor environment. We focus on our evaluation of current stereo techniques. We report the results of the first phase of this evaluation and discuss plans for the second phase. In addition, we briefly describe our progress in the following areas: Development of new object tracking techniques; application of a new scene modeling technique (developed on a parallel contract at SRI) to UGV tasks; and development of two auxiliary techniques to support our research: a technique for generating synthetic stereo pairs and a technique for multiplexing imagery from a pair of moving cameras onto a single videotape.

14. SUBJECT TERMS

Stereo analysis, evaluation of stereo, scene geometry, unmanned ground vehicle, computer vision

15. NUMBER OF PAGES

33

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNLIMITED

TABLE OF CONTENTS

LIST OF FIGURES

PREFACE

This research is sponsored by the Advanced Research Projects Agency (ARPA), 1400 Wilson Boulevard, Arlington, Virginia 22209-2308 and monitored by the U.S. Army Topographic Engineering Center (TEC), Fort Belvoir, Virginia 22060-5546, under Contract DACA76-92-C-0003, by SRI International, Menlo Park, CA 94025-3493. The following SRI International researchers have contributed to the work described in this report: H.H. Baker, R.C. Bolles, M.A. Fischler, P. Fua, M.J. Hannah, J. Herson, and J. Woodfill. The Contracting Officer's Representative was Ms. Linda Graff.

1 Introduction

Starting with the earliest work in computer vision, stereo analysis has played a central role in scene modeling systems. Yet, to date, there has been no comprehensive characterization of its strengths and weaknesses. To remedy this situation, SRI, JPL, and Teleos have initiated a multiphase process for evaluating existing stereo techniques. Our goal is to develop analytic, behavioral, and statistical models of the effectiveness of stereo applied to Unmanned Ground Vehicle (UGV) tasks. In this report, we briefly describe the results of our initial evaluation, which we called the JISCT Stereo Evaluation after the five groups who contributed imagery: JPL, INRIA (in France), SRI, CMU, and Teleos.

In addition, we briefly describe progress in using multiple images to obtain an integrated geometric and physical description of a scene. We present two techniques for tracking moving objects and discuss how a technique developed on a related SRI project can be used to produce robust three-dimensional descriptions of a scene.

2 Evaluation of Current Stereo Techniques

Stereo analysis, which for a long time had been viewed as an interesting, but too-costly-to-be-practical technique, has emerged as a viable tool for realtime applications, such as vehicle navigation. This has happened for two reasons. First, advances in hardware have made it practical to compute stereo matches "in real time." And second, advances in algorithm development have made it possible to correctly match large portions of outdoor scenes.

An important next step in the development and use of practical stereo systems is the characterization of their capabilities. Potential users, such as system integrators and automatic task-planners, need to know the techniques' computational requirements, their speeds, the precision of their results, their common mistakes, and so forth, in order to model the behavior of these stereo systems and reason about their use. With this in mind, SRI, JPL, and Teleos began a multiphase evaluation process last year within the ARPA Unmanned Ground Vehicle (UGV) Project. The first phase of that evaluation has been completed, and the second phase has begun.

The overall plan for that evaluation was (and continues to be) to pursue a three-pronged approach, including analytic models, qualitative "behavioral" models, and statistical performance models. The analytic models would be used to estimate such things as the expected depth precision computable with a specific camera configuration. The qualitative models would be used to identify key problems for future research, for example, detection of holes, analysis of shadowed regions, and depth measurements in bland areas. The statistical models would be used to produce quantitative estimates of such key factors as the smallest obstacle detectable at a specified

distance. SRI has taken the lead in the qualitative evaluation; JPL has taken the lead in the quantitative analysis.

For the qualitative analysis we decided to start by examining a small number of techniques in order to debug the process, and then expand the evaluation to include a much larger set of participants. The goals of the first phase were to get an initial estimate of the effectiveness of current stereo techniques applied to UGV tasks, to identify key problems for future research, and to debug the evaluation process.

One of the high-level guidelines we adopted was to develop and maintain an atmosphere of cooperation and constructive criticism among the researchers participating in the evaluation. Without this we would not be able to focus on our ultimate goal of producing a sequence of increasingly capable stereo systems. To help establish a cooperative atmosphere, we decided to concentrate on the positive aspects of each technique and emphasize potential extensions, realizing that existing techniques were developed for different domains and different applications. We also decided to share all the raw results with the participants so they could duplicate our analysis or develop their own.

For the first phase of the qualitative evaluation, SRI collected imagery from five groups, JPL, INRIA (in France), SRI, CMU, and Teleos (hence the name "JISCT" for the first evaluation phase); selected 49 image pairs for analysis; converted them into a standard format; distributed the dataset to the five groups for processing, along with an extensive set of instructions; collected the results; characterized them; and finally distributed the results and the associated report to the participants.

We intentionally asked each group to process a large number of pairs (10 training pairs and 45 "test" pairs ... 6 pairs were in both the training and test sets), because we wanted to force each group to establish a standard algorithm that was automatically applied. As a result of this approach, there are now three or four groups around the world that can readily apply end-to-end stereo techniques to new data and compare their results. As part of the second phase we hope to expand this community to 10 or more groups. This process has opened up a new form of interaction within the computer vision community that we feel will help stimulate advances and reduce redundant development.

In the instructions to the participants we asked each group to produce several results for each match point in addition to its computed disparity. For each point we asked for an x and a y disparity, an estimate of the precision associated with each reported disparity, an estimate of the confidence associated with each match, and an annotation for each unmatched point, indicating why the technique could not find a match. Possible explanations for no match included "area too bland," "multiple choices," and "inconsistent with neighbors." Although none of the groups produced all this additional information (they all produced some of it), we felt that it was important to begin the process with the goal of producing this auxiliary information, which will be invaluable for the higher-level routines using the stereo results. We

foresee a time in the not too distant future when the calling routine will use the precisions, confidences, and annotations to actively control the sensor parameters for the next data acquisition step. For example, if the current stereo results contain a large region with no disparities and the image regions are quite dark, the controlling routine could open the irises or increase the integration time to reexamine these dark regions.

To assist in the analysis of the results, SRI developed two sets of routines, one to gather statistics and one to display the disparities in a variety of ways. Since we did not have ground truth for the distributed imagery, we were not able to compare the computed disparities with objective values. However, we were able to gather statistics on two of the three types of mistakes in which we were interested by outlining selected regions in the imagery and counting the occurrence of results/no-results within these regions. We made a distinction between the following three types of mistakes:

False Negatives: No disparities computed for points that should have results.

False Positives in Unmatchable Regions: Disparities reported for points that don't have matches in the second image, for example, points occluded in one image or points out of the field of view of one of the images.

False Positives in Matchable Regions: Incorrect disparities reported for matchable points.

By interactively outlining regions of occluded points, regions of points out of the field of view of the second image, and regions of points in the sky, we were able to directly measure statistics for the first two types of mistakes. In addition, we outlined regions corresponding to expected problems, such as dark shadows, foliage, and bland areas. In this way we could gather statistics on the behavior of the algorithms on these special problems.

The results of the first-phase evaluation can be summarized as follows:

- We were surprised by the completeness of the results. Even though the dataset contained a wide range of imagery, including some sequences designed to stretch the analysis along specific dimensions, such as noise tolerance and disparity range, the stereo systems computed disparities for 64% of the matchable points. On eight image pairs selected to be the most appropriate for UGV applications, the techniques computed disparities for as much as 87% of the points. Although the missing points (and mistakes in the reported matches) could cause problems for vehicle navigation, this level of completeness is an indication that there is a solid basis for building a passive ranging system for an outdoor vehicle.

- For the UGV-related imagery the number of gross errors was relatively small, ranging from a few "spike" errors to small regions of mistakes. We estimate that these results contained gross errors of somewhere between 1 and 5%. Many of these errors would have to be eliminated in order for the data to be used directly for planning navigable routes.

- The stereo systems made different mistakes, most of which could be explained by their correlation patch size, search technique, or match verification technique. However, since they made different mistakes, there is a possibility of combining them in a way to check each other and fill in missing data.

- All the stereo systems could be improved significantly with a relatively small amount of effort. This was the first test of this type, requiring the analysis of a large dataset, and it uncovered some weaknesses in the different stereo systems that can be corrected. One area to be considered is the development of preanalysis techniques to automatically set key parameters, such as patch size and search areas (as Teleos did). The filtering of results could also be improved, eliminating matches that differ significantly from their neighbors (as SRI did).

- There were a few surprises, such as Teleos's successful solution to one set of image pairs from CMU that includes a carpet with a repetitive pattern on it. Teleos's large patches were able to detect large regions of subtle differences, which allowed recovery of the correct disparities.

Additional information about the JISCT evaluation, its results, and our goals for the second phase, can be found in Appendix A. [Bolles, Baker, & Hannah].

3 Moving Object Detection, Tracking, and Recognition

Our high-level goal for this research effort is to develop automated methods for producing three-dimensional models of scenes containing moving objects. Our approach is to analyze sequences of temporally coherent images, because they provide the machine with both "redundant" information and new information about the scene. The redundant information can be used to increase the precision and reliability of computed models; the new information can be used to extend models into previously unseen areas. Recently we have developed two new techniques of this type. One is a real-time technique designed to provide feedback within an "active vision" paradigm. The other integrates object recognition into the tracking process in order to bridge gaps in tracking continuity caused by such things as occlusion and low-level processing mistakes.

The first technique, which is the product of a joint effort between Xerox PARC, Stanford University, and SRI, produces motion results (or stereo disparities) at 10 to 15 hertz. It has been implemented on two multi-processor configurations, a 16k-processor Connection Machine and a 5-processor VX/MVX graphics accelerator system (200 MIPS). With these systems we have demonstrated real-time control of a five degree-of-freedom camera system tracking a person walking around a room [Woodfill].

The second technique incorporates object recognition procedures into the tracking process in order to improve tracking reliability and facilitate object identification. Our strategy has involved four steps. First, we train the system to recognize an object, such as a truck, by showing it to the system from several viewpoints. Second, given an image sequence of the truck moving in front of the camera system, we apply our "weaving-wall" tracking technique [Baker & Garvey] to build a temporal model of the objects in the sequence. Third, we apply the PRS recognition system [Chen & Mulgaonkar] to identify the truck in individual images. And fourth, we use the recognition results to "explain" discontinuities in the weaving wall so that we can produce a more coherent description of the motion in the scene.

4 Geometric Recovery

The goal of geometric recovery is to build a three-dimensional structural description of a scene to support such tasks as robot navigation and cartographic modeling. Ideally the description would consist of several interconnected representations, including a detailed representation of the support surface (i.e., the ground), a list of material types and semantic labels for all scene "objects," and a set of accurate transformations from the local vehicle coordinate system to the global reference system. For many tasks, especially robot navigation, the process of building a scene model should be viewed as an ongoing process in which a continuous stream of data is used for incrementally updating the representations. In practice, however, current scene modeling techniques typically analyze each snapshot of a scene independently and produce a loose patchwork of representations, including such things as ground surface patches, clouds of x-y-z points associated with objects, and a set of imprecise transformations from the local coordinate system to the global system.

Our research goals in this area are to develop compact and expressive representations for modeling natural objects, such as rocks and bushes, and to develop effective techniques for incrementally compiling a complete scene model from multiple views. As part of a separate, but related, contract at SRI, we are developing a representation scheme for three-dimensional natural objects and a technique for instantiating and refining object models in this scheme. The following briefly describes our progress in this area and indicates how these techniques apply to our UGV stereo effort.

4.1 Three-Dimensional Object Models

In the past we have developed a number of three-dimensional object representations, including fractal-based descriptions [Pentland], contextual representations [Strat & Fischler], and a "representation space" approach [Bobick & Bolles]. Recently we have developed a triangulated mesh model that supports both object segmentation and surface refinement techniques. In the 1992 Image Understanding Workshop Proceedings we described a technique for coalescing clouds of three-dimensional points into a small number of representative surfaces [Fua & Sander]. In the past year we have concentrated on a specialization of the triangulation representation that we call hexagonally-connected triangular meshes. These meshes have the advantage that they can be easily deformed to refine their local shape so that they satisfy both photometric and depth constraints. The use of these meshes as part of a technique for integrating stereo processing and photometric analysis is presented in a separate paper in these proceedings [Fua & Leclerc].

One advantage of a triangular mesh representation is that many computers now incorporate special hardware to support and perform graphic operations on such representations. This same hardware can be used, with appropriate analysis routines, to predict such things as scene depth values, surface orientations, and observed intensities. We have implemented some of our techniques on Silicon Graphics computers that support these operations.

Since these mesh representations are three-dimensional, they can directly encode all aspects of an object's appearance in a single structure. This structure, in conjunction with rendering techniques, provides a convenient way to work with complex, convoluted objects.

In most of our experiments, we have used regular meshes. While this is appropriate for surfaces whose properties remain relatively constant, it is not optimal for complex surfaces that require the combined efficiency and accuracy provided by irregular networks. The relatively smooth parts of such surfaces can be represented by large patches, while the rougher parts could be described by finer, more precise triangulations. We are in the process of implementing irregular networks formed by allowing selected facets to be subdivided.

4.2 Integration of Stereo and Photometric Analysis

Over the past few years we have investigated techniques for integrating stereo and photometric analysis because these two techniques are complementary; one works well when the imagery contains distinctive photometric patterns, and the other works well when the imagery contains only gradual shading. In 1991 we reported the results of our first technique of this type [Leclerc & Bobick]. Recently we have developed a new approach that functions well even though we have relaxed several assumptions

commonly used in shape-from-shading techniques. This new technique computes both the shape and reflectance properties of physical surfaces from the information present in multiple images. It considers two classes of information. The first is the information that can be extracted from a single image, such as texture gradients, shading, and occlusion edges. The technique takes advantage of the fact that multiple images enhance the utility of this type of information by allowing both consistency checks to filter out mistakes and averaging to improve precision. The second class of information includes the stereo depth values computed from two or more images.

Our surface reconstruction method uses an object-centered representation, specifically a hexagonally-connected three-dimensional mesh of vertices with triangular facets. Such a representation accommodates the two classes of information mentioned above, as well as multiple images (including motion sequences of a rigid object) and self-occlusions. We have chosen to model the surface material using a Lambertian reflectance model with variable albedo, though generalizations to specular surfaces are possible. Consequently, the natural choice for the monocular information source is shading, while intensity is the natural choice for the image feature used in multi-image correspondence. Not only are these the natural choices when we are able to assume a Lambertian reflectance model, they are complementary: intensity correlation is most accurate wherever the input images are highly textured, and shading is most accurate when the input images have smooth intensity variation. Since we wish to deal with surfaces with nonuniform albedo, we have developed a new approach that analyzes the facet-to-facet geometry and albedo pattern to recover surface models.

We use an optimization approach to reconstruct the surface shape and its albedo from the input images. We alter the shape and reflectance properties of the surface mesh to minimize an objective function, given an initial surface estimate provided by other means, such as a standard stereo algorithm. The objective function is a linear combination of an intensity correlation component, an albedo variation component, and a surface smoothness component. The first two components are a function of the intensities projected onto the triangular facets from the input images (taking occlusions into account), and are weighted according to the amount of texture in the intensities, for the reasons mentioned in the previous paragraph. The geometric smoothness component is slowly decreased during the optimization process to allow for an accurate final estimate of the surface shape and reflectance.

We have implemented an algorithm employing these three terms and have performed extensive experiments using synthetic images as well as aerial and face images. The strengths of the approach include:

- The use of the three-dimensional surface mesh allows us to deal with self-occlusions and thus effectively merge information from several potentially very different viewpoints to eliminate "blind-spots."

- By combining stereo and shape from shading, and weighing appropriately the reliability of their respective contributions, we can obtain results that are better than those produced by either technique alone.
- Using the facets to perform the stereo computation frees us from the constant-depth assumption that standard correlation-based stereo techniques make. It becomes possible to recover accurately the depth of sharply sloping surfaces (such as that of a sharp ridge).
- The shape-from-shading component does not make the constant-albedo assumption common to most shading algorithms. Instead, we only make the weaker and much more general assumption that albedoes vary slowly across textureless areas.

More complete details of this technique will appear in the 1993 ARPA Image Understanding Workshop [Fua & Leclerc].

5 Auxiliary Techniques

We have developed two techniques to support our ongoing research efforts. The first one is a technique to generate more realistic synthetic stereo pairs than previously available, and the second is a method for interleaving images taken from a pair of moving cameras onto a single videotape.

5.1 Synthetic Stereo Pairs

We are developing a new interactive technique for generating synthetic stereo pairs, which we expect to play an important role in future stereo evaluations, because complete ground truth will be known for these images. The idea is to compute disparities for a real image pair, interactively correct mistakes, and then use the refined disparity image as the ground truth for constructing a new image pair. Figures 1 through 6 show an example of this process. Figure 1 is the original pair taken at Stanford University. (The left image is on top.) Figure 2 shows the disparities computed from one of our stereo techniques. The results contain several "mistakes." There are regions with no disparity results (shown in black); There are a few gross errors, such as the bright points near the lower left corner; And the closest tree is considerably wider than it should be. To correct these mistakes, a person interactively outlines the occlusion edges in the original left image (shown in Figure 3) and eliminates gross errors. The program then smoothes the data, avoiding the occlusion edges, and fills in missing regions to produce a complete image of disparities, shown in Figure 4. A relative camera model is then used to convert the disparities in Figure 4 into a

three-dimensional scene model. (We use a hexagonally-connected triangular mesh representation for the scene model, just as Fua and Leclerc used in their technique for integrating stereo and photometric analysis [Fua & Leclerc].) Finally, we create a new pair of images of this scene by texture mapping the original left image onto images created for cameras located at the same positions with respect to the scene as the original cameras were. Figure 5 shows this new pair of images. The right image (on the bottom) has a large blank region on the right because that part of the scene is not visible in the left image. To fill in that region, and any other occluded regions, we extract the corresponding parts from the original right image and insert them. Figure 6 shows the final pair of synthetic images.

The result of this process is a new pair of images that look almost identical to the original pair, and the disparities and occlusion regions are known. Pairs of this type will help in our future evaluations of stereo matching techniques, making it possible to compute disparity errors across the whole image. However, this type of imagery will not replace real imagery for at least two reasons. First, the image generation process introduces subtle differences at occlusions, which are important test regions. And second, the process introduces a bias to the ground truth measurements because a stereo technique was used to compute it. As a result, we expect that stereo technique to produce better results on the new imagery than other techniques. And, in fact, we expect that all stereo techniques to do better on the synthetic pair than on the original pair, because of the stereo analysis and smoothing used to create the synthetic imagery.

5.2 Stereo Multiplexing

We have developed a way to multiplex the video from two synchronized cameras onto one videotape by interleaving the even fields from both cameras. The new videotape contains pairs of images taken simultaneously, but stored sequentially on the tape. We used a Datacube MV20 system, a digicolor digitizer, and a maxWare-like set of routines to do this. We input the left camera's data on the RED channel of the digicolor, the right camera's data on the GREEN channel, store both even and odd fields from both channels, ignore the odd fields, and set up the output registers to write the even field of the left camera as the even field of the new video stream and the even field of the right camera as the odd field. This sequence of video fields is then sent back through the digicolor board to convert it back to analog for storage on a videotape.

Recently, the three stereo contractors in the UGV Program – JPL, Teleos, and SRI – agreed on this alternating field format as the standard way to store sequences of stereo pairs on tape. It provides a convenient way to store long sequences of data and/or results that can be easily viewed in stereo via LCD-shuttered glasses.

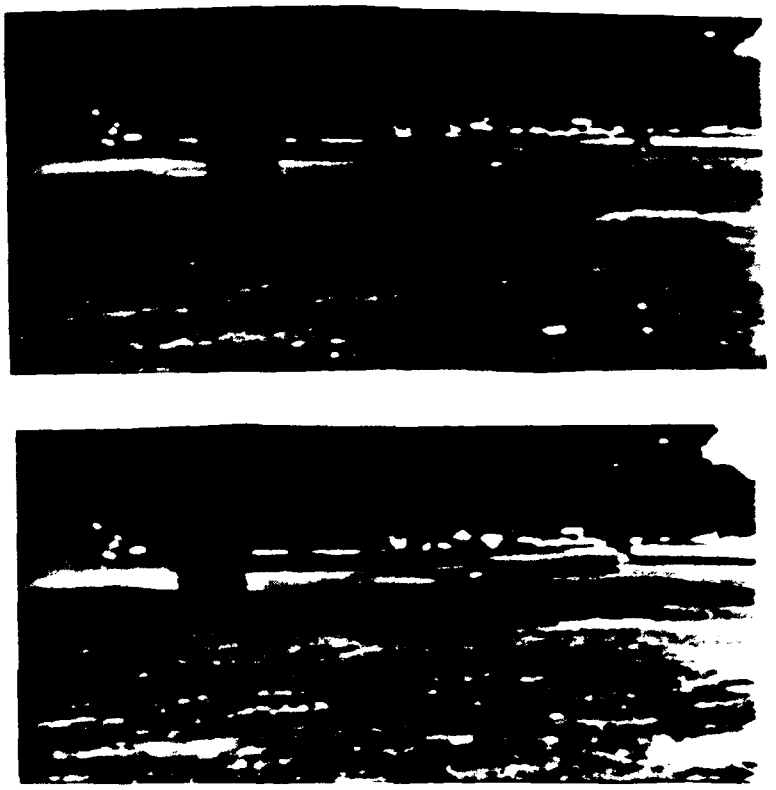


Figure 1: Real stereo pair (left image on top).

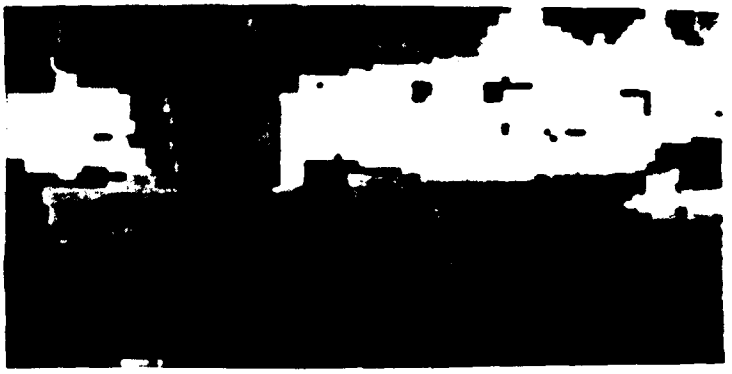


Figure 2: Disparities computed for the pair in Figure 1 (aligned with the left image).

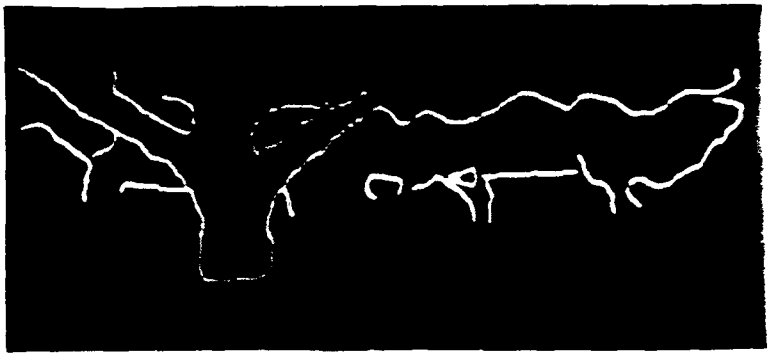


Figure 3: Occlusion edges interactively drawn on the left image in Figure 1.

Figure 4: Interpolated and smoothed disparities.

Figure 5: Preliminary synthetic pair corresponding to the pair in Figure 1.

Figure 6: Final synthetic pair corresponding to the pair in Figure 1.

The previous technique for multiplexing the output from two cameras was to store the even field from the left camera and then the odd field from the right camera. The problem with this method is that the two fields are not taken at the same time. They are taken a 1/60th of a second apart. If the pair of cameras is mounted on a vehicle moving 30 Km/h, the vehicle moves approximately 15 cm between the even and odd fields, which is a significant change in the stereo camera configuration. Since the amount of movement between cameras depends on the speed of the vehicle and the smoothness of the road, the stereo system would have to recompute the relative camera configuration for each stereo pair. To avoid this problem, we have decided to standardize on a convention in which we interleave fields taken simultaneously. It is more difficult to produce videotapes in this form, but the analysis of the data is more straightforward.

6 Summary

We have made significant progress in achieving our goal of evaluating and advancing the state of the art in passive geometric recovery for UGV applications by:

1. Establishing a network of the world's foremost research centers concerned with robotic vision to evaluate current technology and share new advances;
2. Carrying out the first phase of the stereo evaluation task to quantify performance and identify critical problem areas. This work involved the design of new techniques for data acquisition and performance analysis;
3. Developing new techniques for bridging the gap between the raw depth arrays produced by current stereo techniques and the more integrated geometric models required by the UGV control systems.

7 Future Plans

Our plans for the coming year are to (1) continue our development of techniques for building robust models of natural scenes to support UGV navigation (by integrating information from several images), (2) develop techniques for detecting holes and ditches and other navigational hazards in sequences of stereo imagery, and (3) perform a second phase of stereo evaluation.

8 Bibliography

Baker, H.H., and T.D. Garvey, "Motion Tracking on the Spatiotemporal Surface," *Proc. Motion Vision Workshop,* IEEE Press, Princeton, New Jersey, 340–345, October 1991.

Bobick, A.F., and R.C. Bolles, "The Representation Space Paradigm of Concurrent Evolving Object Descriptions," *IEEE PAMI*, Vol. 14, No. 2, pp. 146-156, February 1992.

Bolles, R.C., H.H. Baker, and M.J. Hannah, *in Appendix A.*

Chen, C-H., and P. G. Mulgaonkar, "Uncertainty Update and Dynamic Search Window for Model-Based Object Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* Maui, Hawaii, 692–694, June 1991.

Fua, P.V., and P. Sander, "Reconstructing Surfaces from Unstructured 3D Points," *Proc. DARPA Image Understanding Workshop*, San Diego, California, January 1992.

Fua, P.V., and Y. Leclerc, "Object-Centered Surface Rconstruction: Combining Multi-Image Stereo and Shading," *Proc. ARPA Image Understanding Workshop*, Washington, D.C., April 1993.

Leclerc, Y.G., and A.F. Bobick, "The Direct Computation of Height from Shading," *Proc. 1991 Computer Society Conference on Computer Vision and Pattern Recognition*, Lahaina, Maui, Hawaii, June 1991, (also in *Proc. DARPA Image Understanding Workshop*, San Diego, California, January 1992).

Pentland, A.P., "Fractal-Based Description of Natural Scenes," *IEEE PAMI* Vol.6, No.6, pp. 661-674, 1984.

Strat, T.M., and M.A. Fischler, "Natural Object Recognition: A Theoretical Framework and Its Implementation," *Proc. IJCAI-91*, Sydney, Australia, August 1991.

Woodfill, J., "Motion Vision and Tracking for Robots in Unstructured Environments," Ph.D. thesis, Computer Science Department, Stanford University, August 1992.

Appendix

"The JISCT Stereo Evaluation"
R.C. Bolles, H.H. Baker, and M.J. Hannah

in the Proceedings of the DARPA
Image Understanding Workshop
Washington, D.C., April 1993

The JISCT Stereo Evaluation*

Robert C. Bolles, H. Harlyn Baker, and Marsha Jo Hannah

Artificial Intelligence Center, SRI International
333 Ravenswood Ave., Menlo Park, CA 94025
(bolles@ai.sri.com baker@ai.sri.com hannah@ai.sri.com)

Abstract

The results of the "JISCT" Stereo Evaluation (named after the five groups contributing imagery: JPL, INRIA (in France), SRI, CMU, and Teleos) are presented. The goals of this evaluation, which was the first phase of a multiphase evaluation process, were (1) to get an initial estimate of the effectiveness of current stereo techniques applied to Unmanned Ground Vehicle (UGV) tasks, (2) to identify key problems for future research, and (3) to debug the evaluation process so that it can be repeated with a larger group of participants. SRI collected 49 pairs of images, distributed them to the five participants, and received complete results from three groups — INRIA, SRI, and Teleos. SRI compared the results by interactively analyzing them and automatically gathering statistics.

We were surprised by the completeness of everyone's results. On the eight image pairs that we thought were the most representative of UGV tasks, the techniques computed disparities for as much as 87% of the points with only a few "spike" errors and some scattered regions of points without matches. Although the missing points (and mistakes in the reported matches) could cause problems for vehicle navigation, this level of completeness is an indication that there is a solid basis for building a passive ranging system for an outdoor vehicle. On the other hand, none of these techniques have "solved the stereo problem" — we selected a number of important areas for future research, including filtering out gross errors and handling the wide dynamic range of intensities common in outdoor imagery.

1 Introduction

Stereo analysis, which for a long time had been viewed as an interesting, but too-costly-to-be-practical technique, has emerged as a viable tool for realtime applications such as vehicle navigation. This has happened for two reasons. First, advances in hardware have made it practical to compute stereo matches "in real time." And second, advances in algorithm development have made it possible to correctly match large portions of outdoor scenes.

An important next step in the development and use of practical stereo systems is the characterization of their capabilities. Potential users, such as system integrators and automatic task planners, need to know their computational requirements, their speeds, their precision, their mistakes, and so forth, in order to model their behavior and reason about their use. With this in mind, SRI, JPL, and Teleos began a multiphase evaluation process last year within the Unmanned Ground Vehicle (UGV) Project. The first phase of that evaluation has been completed, and the second phase has begun. This paper describes the results of the first phase.

The overall plan for our complete evaluation process is to pursue a three-pronged approach, including analytic models, qualitative "behavioral" models, and statistical performance models. The analytic models would be used to estimate such things as the expected depth precision computable with a specific camera configuration. The qualitative models would be used to identify key problems for future research, for example, detection of holes, analysis of shadowed regions, and measurement of bland areas. The statistical models would be used to produce quantitative estimates of such key factors as the smallest obstacle detectable at a specified distance. SRI has taken the lead in the qualitative evaluation; JPL has taken the lead in the quantitative analysis.

For the qualitative analysis, we decided to start by examining a small number of techniques in order to debug the process, and then expand the evaluation to include a much larger set of participants. The goals of the first phase were to get an initial estimate of the effectiveness of current stereo techniques applied to UGV tasks and, from this, to identify key problems for future research.

One of the high-level guidelines we adopted was to develop and maintain an atmosphere of cooperation and constructive criticism among the researchers participating in the evaluation. Without this we would not be

*Supported by Advanced Research Projects Agency Contract DACA76-92-C-0003.

able to focus on our ultimate goal of producing a sequence of increasingly capable stereo systems. To help establish a cooperative atmosphere, we decided to concentrate on the positive aspects of each algorithm and highlight ways to strengthen existing techniques, realizing that they were developed for different domains and different applications. We also decided to share all the raw results with the participants so they could duplicate our analysis or develop their own.

For the first phase of the qualitative evaluation, SRI collected imagery from five groups, JPL, INRIA (in France), SRI, CMU, and Teleos (hence the name "JISCT" for the first evaluation phase); selected 49 pairs for analysis; converted them into a standard format; distributed the dataset to the five groups for processing, along with an extensive set of instructions; collected the results; characterized them; and finally distributed the results and the associated report to the participants.

We intentionally asked each group to process a large number of pairs (10 training pairs and 45 "test" pairs ... 6 pairs were in both the training and test sets; we made an administrative mistake on one of the test pairs, reducing the total to 44), because we wanted to force them to establish a standard algorithm that was automatically applied. As a result of this, there are now four groups around the world that can readily apply end-to-end stereo techniques to new data and compare their results. As part of the second phase we hope to expand this community to 10 or more groups. This process is opening up a new form of interaction within the computer vision community that we feel will help stimulate advances and reduce redundant development.

In the instructions to the participants, we asked each group to produce several results for each matched point in addition to its computed disparity. For each point we asked for an x and a y disparity, an estimate of the precision associated with each reported disparity, an estimate of the confidence associated with each match, and an annotation for each unmatched point, indicating why the technique could not find a match. Possible explanations for no match included "area too bland," "multiple choices," and "inconsistent with neighbors." Although none of the groups produced all this additional information (they all produced some of it), we felt that it was important to begin the process with the goal of producing this auxiliary information, which will be invaluable for the higher-level routines using the stereo results. We foresee a time in the not too distant future when the calling routine will use the precisions, confidences, and annotations to actively control the sensor parameters for the next data acquisition step. For example, if the current stereo results contain a large region of points without disparities and the image region is quite dark, the controlling routine could open the irises or increase the integration time to reexamine these dark regions.

Four groups returned results and write-ups to SRI — Teleos, SRI, and two from INRIA. One of the INRIA sets was from a technique that locates linear features and then matches these features. Since this technique reports only disparities along the matched edges, it was not possible to directly compare its results to the others. Therefore, we concentrated our analysis on the three correlation-based algorithms.

Each participating group analyzed its own results. In addition, Harlyn Baker and Marsha Jo Hannah of SRI analyzed the results from all the groups on all 44 pairs and wrote short reviews of them. In the full report [Bolles, Baker, & Hannah], their comments are included as appendices. These comments, plus the automatically compiled statistics, form the core of this evaluation.

Initially, we were a little reluctant to compute and publish statistics that may be taken out of context. On the other hand, statistics, if reported with sufficient caveats, can provide a convenient basis for comparing techniques. In this paper, we summarize the qualitative results and quantitative statistics. The validities of both are limited by the dataset, which implicitly defines the range of data for which the conclusions directly apply, and by the analyzers, who naturally focused on issues they were most interested in.

This paper is organized as follows. In Section 2, we briefly describe the key strategies and parameters of the three principal techniques, highlighting their similarities and differences. In Section 3, we describe our experimental procedure. In Section 4, we present the automatically gathered statistics, which we refer to as the believe-everything-they-tell-you statistics because they are based on the number of "reported" disparities in specified regions of the test data, not on the number of "correct" disparities. In Section 5, we summarize our qualitative analysis and briefly discuss open issues for future research. In Section 6, we conclude with an evaluation of the JISCT evaluation and make some suggestions for the next step in the evaluation process.

2 Technique Summaries

We evaluated three techniques, whose key aspects are highlighted below.

2.1 INRIA

This technique was originally implemented as part of a European space project to produce three-dimensional models of scenes containing rocks and sand. It is implemented in C on a Sun. A similar technique is implemented on a Connection Machine (by Pascal Fua) at SRI. Key aspects are

- The algorithm computes a disparity for every pixel in an image by matching patches (usually 11x11 pixels) at one or two image resolutions, independently.

The basic algorithm "INRIA-1" matches only at one resolution.

- The technique uses an approximation to normalized correlation, referred to as C5, because it can be implemented efficiently using a sliding computation of the basic sums.
- The algorithm searches only along epipolar lines, which are assumed to be horizontal.
- The algorithm expects a range of disparities to be specified for each image pair to be analyzed.
- The technique verifies all matches by independently matching patches from the left image in the right image and patches from the right image in the left image. If the match for a patch from the left image is not mapped back to within a pixel of its location in the left image, the point is not assigned a disparity.
- The technique computes a subpixel location for each match by fitting a second-order curve to the correlation values surrounding the best match.
- After computing disparities for as many pixels in the lef. image as possible, the algorithm filters out isolated matches by morphologically shrinking the regions of matches. It typically shrinks the regions three times, grows the result three times, and then ANDs this result with the original image of results. This process can erase regions as large as 6x6 pixels.
- The algorithm computes a confidence value for each disparity by differencing the heights of the two highest matching peaks.
- The technique estimates the precision of a disparity value by fitting a Gaussian to the matching peak, using its standard deviation as the precision measure.
- The technique does not attempt matches near the edges of an image.
- The second set of results provided for this evaluation often was produced by matching at two image resolutions and picking the highest resolution for which there was a valid match.

2.2 SRI

This stereo system has evolved over 20 years, beginning with early Martian Rover research, migrating into the aerial mapping domain, and now coming back to ground-level analysis. Its goal has been to produce a *set of high-quality matches from a wide range of (possibly uncalibrated) imagery*. The algorithm is a multistage process that uses one matching technique to get a few solid matches at high-information points, and then uses these matches to guide another matching technique, whose results become anchors for yet another technique, etc, with culling of mistakes occurring at many levels. At each stage, the algorithm acquires more supporting matches to suggest limits for the disparity search, so the algorithm can attempt to match points that have less "interesting" information, using less hierarchy. For this evaluation, code was added to produce "dense" matches: this included stages that grow regions of matches around previously matched points, and fill in a regular grid of matches. In total, the standard algorithm for this evaluation involved seven stages of matching and three filtering steps. The algorithm is implemented in C on a Sun: speed has not been a priority.

Some key aspects are

- The algorithm applies a version of hierarchical matching for each point that it analyzes. At the early stages of the process, it uses all available image resolutions, starting at the coarsest, using the match found at that level to predict the location of the match at the next finer level, then refining it, and so forth. At the final stage, where the dense grid of points is computed, the algorithm uses only one or two levels.
- At each image resolution (level), the algorithm does a two-dimensional search near the epipolar line and then hill-climbs around the best match. The epipolar lines can be at any angle in the second image, and if there is no camera model (due to bad matches at early stages, or because the camera isn't modelable by a pinhole camera), the algorithms search over areas—(dx,dy) boxes—defined by surrounding matches.
- The algorithm uses normalized cross correlation (correcting for a linear intensity change from image to image) on 11x11 patches typically. Later stages, such as the region-growing step, can use smaller patches. The final match includes a subpixel estimate of the disparity, computed by fitting two parabolas to the nearby correlation values.
- Each match from one image to another is verified by applying the same technique to match back into the original image. If the return match is not within a pixel of the original point, the match is discarded as unreliable.
- The algorithm applies several other "filters" to weed out mistakes, including a threshold on interest value, thresholds on relative and absolute correlation values, tests for matches outside an image, and tests for unusual disparity values within a region of the image.
- Later stages of the algorithm use previously computed disparities in the neighborhood of a new point

to be matched, to specify the range of disparities to be considered. The neighborhoods are typically large, beginning at 1/4th of the image area, and gradually reducing to 1/64th of the image for this experiment. This technique assumes that the scene is composed of relatively large continuous surfaces.

- Since a confidence for each match was requested for this experiment, one was supplied by computing the ratio of the correlation value to the autocorrelation threshold.

2.3 TELEOS

This technique has been designed for efficient implementation and recently has been geared toward active vision in which the basic stereo process matches 100 to 200 selected points in a 1/30th of a second. It is implemented on a combination of two special boards and a Datacube system. For this evaluation, however, the hardware was not available and so a Lisp version of the algorithm (running on a Lisp Machine) was used. Some key aspects are

- The algorithm uses large correlation windows (ranging from 24x24 to 96x96 pixels).
- The algorithm computes binary correlation values from the Laplacian of Gaussian of the original images.
- The algorithm analyzes the data only at one resolution. It automatically selects the size of the convolution operators by analyzing the peak shapes of matches at 25 points in each new image pair. It selects the smallest window size that produces a significant difference between the heights of the top two highest peaks.
- At each point in the image, the algorithm starts with the disparity computed for the neighboring pixel and tries to locate a match at a similar disparity. A serpentine search, which analyzes the first row from left to right, the second row from right to left, and so forth, is used in order to reduce the computation time on the Lisp Machine.
- The algorithm searches off the epipolar line for the best match.
- The algorithm also examines the effect of skewing the patch being matched. It analyzes skews ranging from -.5 pixels per line to +.5 pixels per line. This analysis is applied only at the end of the search when the best match has been selected.
- The algorithm estimates a subpixel disparity value by fitting a quadratic function to the best peak.
- The algorithm does not try to match points near the edges of an image.

3 Experimental Procedure

The goal of this initial evaluation was to produce a qualitative characterization of the capabilities of current stereo techniques applied to UGV tasks. The intent, as stated in the instructions distributed to each participant, was to produce a description such as the following:

> On the 44 image pairs in the database our techniques correctly measured disparities to 65% of the points on the ground and 40% of the points on obstacles, such as trees, bushes, and rocks. The top five problems for our techniques were dynamic range, holes, bland areas, repeated structure, and poor range resolution. We estimate that these problems occur in the UGV scenarios with frequencies of ...

The idea was to produce a characterization that would focus future work on key UGV problems.

Our basic approach to developing this type of characterization was to apply the techniques to a large dataset, visually display the results in ways to highlight unusual events, gather basic statistics, and where possible, summarize our observations in descriptions that link observed behaviors to aspects of the techniques.

To start the process, SRI compiled a database of 49 image pairs from JPL, INRIA, SRI, CMU, and Teleos. We converted the images into a standard format and then distributed them to the five contributing groups for analysis. The groups were instructed to use 10 pairs as a training set, "freeze" their algorithm, and then process the whole set of 45 pairs. Results and commentary from four stereo systems were returned to SRI -- Teleos, SRI, and two from INRIA. One of the INRIA sets, using edge-based feature analysis, could not easily be compared with the others. We concentrated our analysis on the three correlation-based system results.

To assist in the analysis of the results, SRI developed two sets of routines, one to gather statistics and one to display the disparities in a variety of ways. Since we did not have ground truth for the distributed imagery, we were not able to compare the computed disparities with objective values. However, we were able to gather statistics on two of the three types of mistakes that we are interested in by outlining special regions in the imagery and counting the occurrence of results within these regions.

We made a distinction between the following three types of mistakes:

False Negatives: No disparities computed for points that should have results.

False Positives in Unmatchable Regions: Disparities reported for points that don't have matches in the second image, for example, points occluded in one

image or points out of the field of view of one of the images.

False Positives in Matchable Regions: Incorrect disparities reported for matchable points.

By interactively outlining regions of occluded points, regions of points out of the field of view of the second image, and regions of points in the sky, we were able to directly measure statistics for the first two types of mistakes. In addition, we outlined regions corresponding to expected problems, such as dark shadows, foliage, and bland areas. In this way we could gather statistics on the behavior of the algorithms on these special problems.

As part of the initial instructions we asked each group to extend its algorithm to produce an image of annotations that summarizes the result of the analysis, pixel by pixel. At each pixel we asked for a code from the following list:

```
 0:  no match attempted
 1:  matched fine

   NO MATCH BECAUSE
 2:  too bland, no information to key on
 3:  low match value (e.g., correlation value)
 4:  multiple choices (ie, repeated structure)
 5:  back-match inconsistency
 6:  point out of camera's field of view
 7:  point occluded by an object in the scene
 8:  point too far off the epipolar line
 9:  point inconsistent with neighbors
10:  other
```

The reason for requesting these codes is to encourage future algorithms to provide this additional information, which can be used by the higher-level vision techniques to decide what should be done next. For example, if no results are reported for a region directly ahead of the vehicle and the region is too bland and very dark, one option might be to open the irises on the cameras (or increase the integration time) in order to see into the dark area.

INRIA reported codes of 1 and 10; SRI reported all codes except for 4 and 7; and Teleos reported codes of 0, 1, 2, and 3. Therefore, we were able to count the number of matches attempted in each region and the number of disparities reported.

To estimate the frequency of incorrectly reported disparities (the third type of mistake), we either compared them to interactively selected values or located an aberration in the local pattern of disparities when they were displayed on the screen. We experimented with a variety of display techniques, including displaying the disparities as color-coded dots in stereo, heights above a three-dimensional "ground" plane, and disparity-displaced vertical lines. We are continuing to look for better ways to display three-dimensional results, because most current techniques encourage the human eye to "smooth over" differences, making the results look better than they actually are.

4 Statistics Summary

The statistics that we refer to as believe-everything-they-tell-you statistics are based on the number of reported disparities in specified regions of the test data. These statistics do *not* distinguish between "correct" and "incorrect" disparity values, just reported values and unreported values. They do, however, provide enough information to estimate three important quantities, the number of false negatives (matchable points that were not assigned a disparity), the number of false positives occurring in unmatchable regions, and the number of matchable pixels that were assigned disparities.

To help focus attention of key areas of the test data, we interactively outlined regions in the left images of 20 of the 44 image pairs (see Figure 1 and Figure 2). One of the most important regions is what we called "matchable-data." It eliminates several types of points that do not have matches in the right image, including null bands that do not contain grayscale data (but are included in the images to fill them out to a standard size, such as 512 by 512 pixels) and pixels that are out of the field of view of the right camera. In the 20 images we examined, the percentage of unmatchable points ranged from 4.3% to 46.0% and averaged 12.3%.

The statistics were gathered by a program that counted the number of disparities (dx disparities) reported in the specified region (or the whole image, if that was appropriate).

Figure 3 shows the results on all 44 image pairs. Note that

- The dataset contains a wide variety of imagery: some of it is realistic (containing dirt roads and cross-country scenes) and some is designed to test the algorithms along one dimension, such as baseline and noise tolerance. Some of the imagery is even trick imagery (the shoe images from CMU).

- The numbers in parentheses after each group's name (along the top of the table) indicate the number of test pairs in the dataset from that group.

- The INRIA-2 results are in parentheses because different parameter settings were used for different image pairs. However, the usual change was for the technique to match at two spatial resolutions instead of just one, and then combine the results. If a second set of parameters was not tried for a pair, we left the entry blank and used the INRIA-1 results in our computation of INRIA-2's average

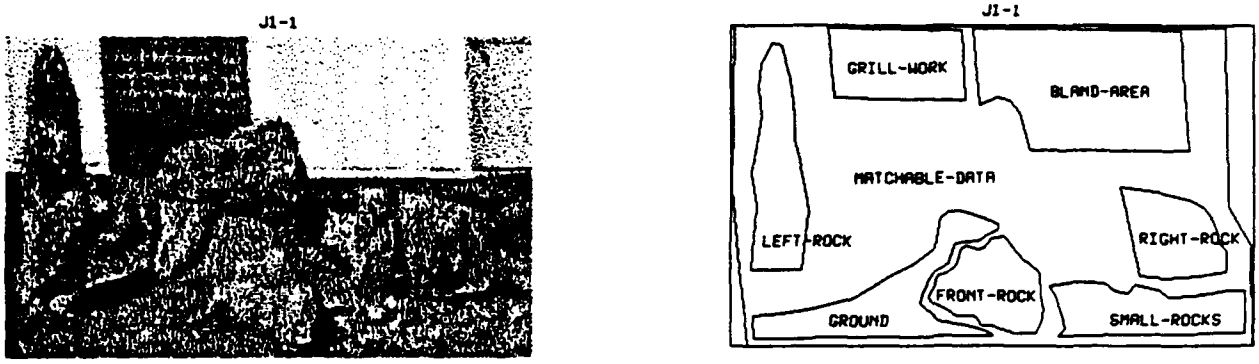


Figure 1: Interactively outlined, special-interest regions for the J1 image pair from INRIA.

- If we did not outline a "matchable-data" region for a pair, we used the full-image statistics in our computations. This reduces the effectiveness totals somewhat (possibly by as much as 7%).

Given the diversity of the data, we were pleased with the completeness of the results.

In order to examine the behavior of the techniques on typical UGV imagery, we selected the eight images from the dataset that were the most appropriate for UGV tasks and collected statistics on that subset. Figure 4 shows the results on these data. The INRIA-2, SRI-2, and Teleos-1 techniques performed well, computing disparities for 86 or 87% of the matchable points. Note, however, that these images did not contain difficult obstacles, such as holes, ditches, and small rocks—the obstacles were large rocks, bushes, and trees.

Figure 5 shows the results on the 17 large obstacles in the dataset. The techniques did an excellent job of detecting these objects, which stick up above the ground — they only had a little trouble in shadowed regions on them.

With respect to shadows, the techniques had a significantly harder time computing disparities for points in shadowed regions than in sun-lit regions. Figure 6 shows the results for points in shadows.

The techniques also had trouble with bland regions, as expected. Figure 7 shows the results on these areas. The techniques typically computed results around the edges of the regions — the larger the correlation windows, the more points were computed, because correlation windows naturally extend matches into the interior of bland regions by about half their diameter.

There are several potentially important problem areas that were not covered in this initial dataset, including holes, sand, small- to medium-sized rocks and bushes, reflective surfaces (water or windows), and moving objects. One of our goals for the second phase of this evaluation is to include examples of these problems.

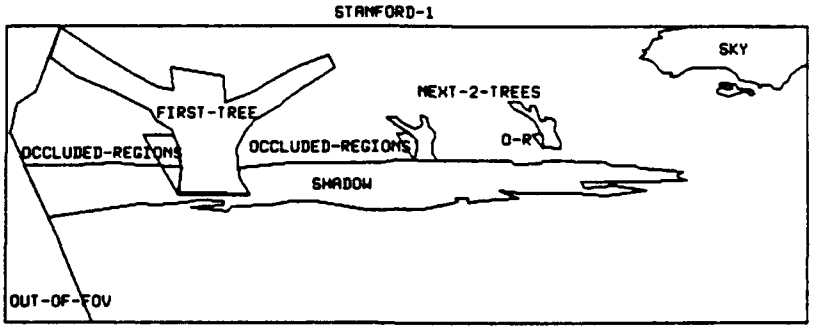
5 Qualitative Analysis

We were surprised by the completeness of everyone's results. Even though the dataset contained a wide range of imagery, including some sequences designed to stretch the analysis along specific dimensions, such as noise tolerance and disparity range, the techniques computed disparities for 64% of the matchable points. On the eight image pairs that we selected as the most appropriate for UGV applications, the techniques computed disparities for as much as 87% of the points. Although the missing points (and mistakes in the reported matches) could cause problems for vehicle navigation, this level of completeness is an indication that there is a solid basis for building a passive ranging system for an outdoor vehicle.

The number of gross errors varied considerably from image pair to image pair. For most "realistic" images the number was relatively small, ranging from a few "spike" errors to small regions of mistakes. We estimate that for these images there were between 1 and 5% gross errors in the results. In many cases, the worst errors cluster into areas that are "breaking up" for one reason or another (usually poor information plus a poor "guess" for the disparity range); if we can "fix" these areas, then the remaining "spike" errors should be amenable to culling techniques. In any case, most of these errors would have to be eliminated in order for the data to be used directly for planning navigable routes.

The techniques made different mistakes, most of which could be explained by their correlation patch size, search technique, or match verification technique. However, since they made different mistakes, there is a possibility of combining them in a way to check each other and fill in missing data.

All the techniques could be improved significantly with a relatively small amount of effort. This was the first test of this type, requiring the analysis of a large dataset, and it uncovered some weaknesses that can be corrected. One area to be considered is the development of preanalysis



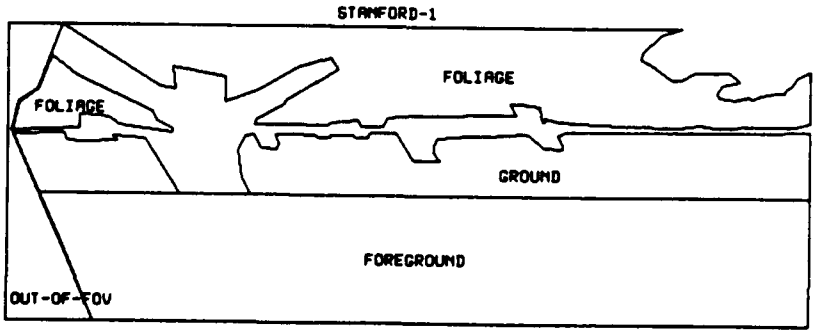


Figure 2: Special-interest regions for the STANFORD image pair from SRI.

	JPL(5)	INRIA(8)	SRI(15)	CMU(9)	Teleos(7)	Weighted Average
INRIA-1	63	66	42	89	35	57
(INRIA-2)	(92)	(75)	(60)	(70)	(50)	(67)
SRI-2	94	74	61	64	39	64
Teleos-1	95	81	45	87	77	71
Average	84	74	49	80	50	64

Figure 3: Percentage of "matchable" pixels assigned disparities on all 44 image pairs.

	Arroyo	EPI16	HMMWV1	HMMWV2	J1	Road	Rock	StanDbl	Average
INRIA-1	90	60	67	79	72	40	37	47	62
(INRIA-2)		(85)	(95)	(95)		(90)	(88)	(76)	(86)
SRI-2	91	72	94	94	73	97	94	78	87
Teleos-1	98	72	93	91	74	98	95	72	87
Average	93	68	85	88	73	78	75	66	79

Figure 4: Percentage of "matchable" pixels assigned disparities on the eight most representative pairs.

	Arroyo			HMMWV1			HMMWV2		Rock		
	Bush1	Bush2	Rock	LMound	Rock	RMound	Rock	Etc	LBush	RBush	Rock
Pixels:	56	68	10	130	18	106	26	839	174	105	23
INRIA-1	95	88	100	98	100	88	100	96	67	30	57
(INRIA-2)				(100)	(100)	(100)	(100)	(98)	(74)	(74)	(100)
SRI-2	91	82	90	99	94	96	96	95	68	64	96
Teleos-1	90	100	90	88	100	97	88	94	74	72	57
Average	92	90	93	95	98	94	95	95	70	55	70

	J1			StanDbl		Ball2	Unweighted
	RRock	FRock	LRock	1Tree	2&3T	Tennis-Ball	Average
Pixels:	70	70	98	276	37	145	
INRIA-1	100	100	100	63	86	92	85
(INRIA-2)				(92)	(100)	(94)	(95)
SRI-2	100	99	99	50	76	90	87
Teleos-1	98	89	100	94	62	86	87
Average	99	96	100	69	75	89	86

Figure 5: Percentage of "matchable" pixels on large obstacles assigned disparities.

	Stanford		StanDbl		Unweighted
	Shadow	1stTree	Shadow	1stTree	Average
Pixels:	215	140	448	276	
INRIA-1	40	71	38	63	53
(INRIA-2)	(84)	(96)	(73)	(92)	(86)
SRI-2	59	61	65	50	59
Teleos-1	1	82	29	94	52
Average	33	71	44	69	55

Figure 6: Percentage of "matchable" pixels in shadows assigned disparities.

	iRoad2 Road	J1 Bland	Ball2 Matchable-	Ball4 Matchable-	Unweighted Average
Pixels:	1077	229	3126	3126	
INRIA-1	12	22	56	39	32
(INRIA-2)	(86)		(72)	(62)	(61)
SRI-2	63	32	76	43	54
Teleos-1	83	10	86	84	66
Average	53	21	73	55	51

Figure 7: Percentage of "matchable" pixels in bland areas assigned disparities.

techniques to automatically set key parameters, such as patch size and search areas (as Teleos does). Another place for improvement is in the filtering of the results to eliminate matches that differ significantly from their neighbors (as SRI and INRIA do).

There were a few surprises, such as Teleos's successful solution to one set of image pairs from CMU that includes a carpet with a repetitive pattern on it. Teleos's large patches were able to detect large regions of subtle differences, which led to the correct disparities.

5.1 Technique-Oriented Summaries

No one of these algorithms has completely solved the stereo problem, although all can produce basically usable results on most reasonable imagery. Each has strengths and weaknesses—and very often an algorithm's strength on one dataset is its weakness on another!

INRIA's algorithms assume that the images are in epipolar alignment. This makes their searches more efficient, and keeps matches from wandering off of the epipolar lines (for instance, "climbing" the edges of tree trunks). However, when presented with nonepipolar imagery, INRIA-1 fell apart; INRIA-2 did better, but had a persistent problem, producing rough disparity contours, which are apparently due to the way the pyramid was handled. The low-resolution results were simply zoomed-out using pixel replication. This epipolar line constraint also limits the usefulness of INRIA's algorithms on imagery from nonpinhole cameras.

SRI's algorithm mostly disregards the epipolar constraint. Consequently, it had no particular problems handling nonepipolar imagery. However, it failed to match many of the very smooth tree edges in the EPI sequence, probably because its matches "slid" up the linear sides of the trees.

INRIA's algorithms search the entire width of the epipolar line. This helped them to do well on some datasets, but when the ground texture was ambiguous, their technique tended to return no match because of multiple choices.

SRI's algorithm depends on early matches to "set the context", so that later searches for matches can be confined to the disparities in that neighborhood. When there is enough global texture for the initial matches to give a good sampling of the disparities, this works well, enabling SRI-2 to produce ground plane matches where the others couldn't. However, when lack of foreground detail keeps SRI-2 from having the right initial matches, it fails to match, or finds random mismatches.

Teleos's algorithm uses very large windows dynamically skewed to accommodate tilted planes. This causes it to do well on some ground planes where it was able to disambiguate the pattern through minor variations, but not on others where the ground plane tilt was out of the allowed range of skewing. Of course, these large windows also cause it to have problems with any scene containing depth discontinuities—it either finds no match, or tries to blend the foreground object into the background objects, or widens the foreground object out onto the background. In addition, Teleos-1's scanning heuristic creates some rather peculiar artifacts— extending objects in opposite directions on alternate scan lines. However, its ability to "see" into low-contrast situations is very good.

The Teleos system, with its large correlation windows, also produces smaller range images, because it limits matching to areas where the full correlation patch is within the image. In an active vision system, the sensors could be reoriented to center objects of interest that may initially appear on the boundary of an image.

Both INRIA's and SRI's algorithms use fairly small windows. This removes much of the need for window skewing and warping, although on extremely tipped planes, warping would be helpful. INRIA-1, INRIA-2, and SRI-2 all do better on tilted planes if the information is slightly "fuzzy". These algorithms don't do nearly as well in the presence of man-made ambiguous patterns.

SRI's algorithm tends to leave more holes in the data—low-information places that it refuses to try to match, ambiguous places where it can't backmatch successfully, or error matches that it has detected and removed. This gives the data a "lacey" appearance, and it should probably be followed by an interpolation step, to fill in these problem areas. (The SRI technique is capable of interpolation, but it was not used in this evaluation.) SRI-2 often leaves a nice band of no-matches outlining depth discontinuities, where one doesn't really want separate objects "smoothed" together. SRI-2 also often refuses to match areas like the sky, which technically don't have a match.

None of the algorithms currently distinguishes between good image data and the "null data" areas caused by image digitization, reprojection, and so forth. This can lead to rather peculiar mismatches around these areas of null data. All of the algorithms should add the ability to accept a mask telling what parts of the image not to try to match. Better yet would be a preprocessing step to construct these masks automatically.

It was interesting to see how much better all of the algorithms did on the imagery taken by JPL than on the SRI imagery. A major factor is the unusual aspect ratio of the SRI imagery caused by digitizing individual fields, since the vehicle was moving fast enough to show a significant difference between fields. JPL's imagery was taken while the vehicle was standing still. Other differences that may have contributed include image contrast, epipolar geometry, and look angle (SRI's cameras were looking far forward, whereas JPL's were looking down a bit more). We note that the exchange of imagery can help in algorithm development by avoiding inadvertently "tuning" one's algorithm to one's particular style of imagery.

5.2 Open Research Problems

After examining the results from this dataset, we have selected the following topics for future research in the area of low-level passive range sensing:

1. Filtering out gross errors caused by erroneous matches.
2. Handling the wide dynamic range in intensities common in outdoor imagery, from dark shadowed regions up to specularities off shiny surfaces.
3. Handling the large range in adjacent disparities arising from narrow foreground obstacles.
4. Adjusting algorithm parameters automatically to properly handle different image regions, such as bland areas and texture regions.
5. Detecting multiple matches and selecting the correct one, possibly by analyzing multiple images.
6. Providing validation and confidence estimation mechanisms.
7. Detecting occlusion edges and reporting accurate depths on both sides of them.
8. Detecting and characterizing small- to medium-sized obstacles, such as rocks and bushes.
9. Detecting "negative" obstacles, such as holes and ditches.

Although the JISCT dataset did not include examples of the last two areas, they are clearly important for cross-country navigation.

6 Conclusion

As a result of this phase of our stereo evaluation, we can make a few general observations and develop a few ideas for the project's next phase.

First, the time is right for evaluation. If promising computer vision techniques, such as stereo analysis and road following, are to make the transition from the research laboratory to practical systems, their characteristics will have to be well enough documented that system engineers can understand them and predict their behavior. We view this evaluation as the first tentative step toward developing this type of characterization.

Second, evaluations of this type require a significant effort. To give an idea of what is involved in such an evaluation, SRI did the following: gathered imagery from five groups, converted it into a standard format, designed the experimental procedure, distributed the imagery to the participants, collected the results, converted them into a uniform format (correcting for a few mistakes in the original specifications), developed visualization routines, used these routines to interactively examine all the results, developed statistics gathering routines, applied these routines to the results, wrote the report, and finally distributed the report and copies of everyone's results.

Third, ideally an evaluation of this type should be performed periodically to provide estimates of the relative improvements of the techniques.

6.1 Critique of the JISCT Evaluation

Some things that were done correctly:

- We developed a cooperative attitude among the participants. This was the first time our community

had tried establishing an ongoing evaluation process and we knew that we'd make mistakes. We also knew that the participants have their egos involved in their systems, and we wanted to emphasize the constructive aspects of comparing techniques.

- The experimental procedure was almost right. The idea of distributing a large number of stereo pairs, using some for a training set, freezing the "official" algorithm, and then applying it to 45 test pairs is correct. The large number of pairs virtually forced the groups to implement an automatic technique, which they could apply to any image pair. As a result, there are now four systems around the world that can be easily tested on new imagery.

- The idea of asking for precision estimates, confidence estimates, and annotations was correct. Although no group produced them all, future systems will be expected to because this information is so important for higher-level users of the results.

- The basic idea of sharing data from several groups was good because applying the algorithms to this diverse set of images brought to light several implicit and explicit assumptions and parameters in the algorithms.

- Since any evaluation of this type can only include a limited set of imagery that attempts to cover all possible dimensions, the idea of including several small controlled experiments worked well. For example, the set of images from Teleos explored the ability of the algorithms to handle increasing noise; the SRI EPI sequence tested a range of baselines.

Some things that should be changed:

- The lack of ground truth significantly limited the types of automatic "objective" evaluations possible. Ground truth is expensive, but there is no substitute for assessing quantitative issues.

- For this initial phase we built our dataset primarily from existing data. In the future we need to gather data that is more realistic and appropriate to the task. In particular, for UGV tasks, the data should be from the demonstration sites and include examples of the common "obstacles," such as ruts, bushes, rocks, ditches, and water. Future datasets should also include sequences of images and trinocular data, not just individual pairs.

- The whole process took too long (almost a year). Techniques can change faster than that. To be relevant, the results should be returned within a few months. This turnaround time is more possible now that we have been through the process once and have developed routines for analyzing the data.

- More auxiliary data (e.g., calibration information) should be supplied with the dataset. Some techniques rely on this information to reduce search and set key parameters. Also, it will generally be available in most applications.

6.2 Plans for the Next Evaluation Phase

We plan to include three types of imagery in the next dataset: demonstration-related pairs and sequences, a few image-intensified pairs, and some synthetic pairs that are less artifactual than previous ones. One of our goals for this phase is to explore more rugged off-road scenes, including deep ruts, tall grass, and ditches, so we are including several examples of each in the new dataset. The image-intensified data will provide our first look at applying our techniques to night-vision-type imagery. The synthetic data is formed from real pairs by modifying a set of computed disparities, and then forming a new right image based on these disparities. This data, although still not completely realistic, is significantly better than previous versions and provides complete ground truth.

We plan to distribute the dataset to 10 or 15 research groups for analysis. After debugging the process, we are in a position to open up the evaluation to include a wider group of participants.

Reference

Bolles, R.C., H.H. Baker, and M.J. Hannah. "The "JISCT" Stereo Evaluation," SRI International Report, January 1993.